The Pipeline Project:
Pre-Publication Independent Replications of a Single Laboratory's Research Pipeline

Martin Schweinsberg	INSEAD
Nikhil Madan	INSEAD
Michelangelo Vianello	University of Padova
S. Amy Sommer	HEC Paris
Jennifer Jordan	University of Groningen
Warren Tierney	INSEAD
Eli Awtrey	University of Washington
Luke (Lei) Zhu	University of Manitoba
Daniel Diermeier	University of Chicago
Justin Heinze	University of Michigan
Malavika Srinivasan	Harvard University
David Tannenbaum	University of Chicago
Eliza Bivolaru	INSEAD
Jason Dana	Yale University
Clintin P. Davis-Stober	University of Missouri
Christilene Du Plessis	Rotterdam School of Management, Erasmus University
Quentin F. Gronau	University of Amsterdam
Andrew C. Hafenbrack	Católica Lisbon School of Business & Economics
Eko Yi Liao	Macau University of Science and Technology
Alexander Ly	University of Amsterdam
Maarten Marsman	University of Amsterdam
Toshio Murase	Roosevelt University
Israr Qureshi	IE Business School
Michael Schaerer	INSEAD
Nico Thornley	INSEAD
Christina M. Tworek	University of Illinois at Urbana-Champaign
Eric-Jan Wagenmakers	University of Amsterdam
Lynn Wong	INSEAD
Tabitha Anderson	Illinois Institute of Technology
Christopher W. Bauman	University of California, Irvine
Wendy L. Bedwell	University of South Florida
Victoria Brescoll	Yale University
Andrew Canavan	Illinois Institute of Technology
Jesse J. Chandler	Mathematica Policy Research; Institute for Social Research, University of Michigan

Erik Cheries	University of Massachusetts Amherst
Sapna Cheryan	University of Washington
Felix Cheung	Michigan State University; University of Hong Kong
Andrei Cimpian	University of Illinois at Urbana-Champaign
Mark Clark	American University in Washington DC
Diana Cordon	Illinois Institute of Technology
Fiery Cushman	Harvard University
Peter H. Ditto	University of California, Irvine
Thomas Donahue	Illinois Institute of Technology
Sarah E. Frick	University of South Florida
Monica Gamez-Djokic	Northwestern University
Rebecca Hofstein Grady	University of California, Irvine
Jesse Graham	University of Southern California
Jun Gu	Monash University
Adam Hahn	Social Cognition Center Cologne, University of Cologne
Brittany E. Hanson	University of Illinois at Chicago
Nicole J. Hartwich	University of Cologne
Kristie Hein	Illinois Institute of Technology
Yoel Inbar	University of Toronto
Lily Jiang	University of Washington
Tehlyr Kellogg	Illinois Institute of Technology
Deanna M. Kennedy	University of Washington Bothell
Nicole Legate	Illinois Institute of Technology
Timo P. Luoma	Social Cognition Center Cologne, University of Cologne
Heidi Maibeucher	Illinois Institute of Technology
Peter Meindl	University of Southern California
Jennifer Miles	University of California, Irvine
Alexandra Mislin	American University in Washington DC
Daniel C. Molden	Northwestern University
Matt Motyl	University of Illinois at Chicago
George Newman	Yale University
Hoai Huong Ngo	Université Paris Ouest Nanterre la Défense
Harvey Packham	University of Hong Kong
Philip S. Ramsay	University of South Florida
Jennifer L. Ray	New York University
Aaron M. Sackett	University of St. Thomas

Anne-Laure Sellier	HEC Paris
Tatiana Sokolova	HEC Paris and University of Michigan
Walter Sowden	University of Michigan
Daniel Storage	University of Illinois at Urbana-Champaign
Xiaomin Sun	Beijing Normal University
Jay J. Van Bavel	New York University
Anthony N. Washburn	University of Illinois at Chicago
Cong Wei	Beijing Normal University
Erik Wetter	Stockholm School of Economics
Carlos Wilson	Illinois Institute of Technology
Sophie-Charlotte Darroux	INSEAD
Eric Luis Uhlmann	INSEAD

Abstract

This crowdsourced project introduces a collaborative approach to improving the reproducibility of scientific research, in which findings are replicated in qualified independent laboratories before (rather than after) they are published. Our goal is to establish a non-adversarial replication process with highly informative final results. To illustrate the Pre-Publication Independent Replication (PPIR) approach, 25 research groups conducted replications of all ten moral judgment effects which the last author and his collaborators had "in the pipeline" as of August 2014. Six findings replicated according to all replication criteria, one finding replicated but with a significantly smaller effect size than the original, one finding replicated consistently in the original culture but not outside of it, and two findings failed to find support. In total, 40% of the original findings failed at least one major replication criterion. Potential ways to implement and incentivize pre-publication independent replication on a large scale are discussed.

Keywords: crowdsourcing science; replication; reproducibility; research transparency; methodology; meta-science

The Pipeline Project: Pre-Publication Independent Replications of a Single Laboratory's Research Pipeline

The reproducibility of key findings distinguishes scientific studies from mere anecdotes. However, for reasons that are not yet fully understood, many, if not most, published studies across many scientific domains are not easily replicated by independent laboratories (Begley & Ellis, 2012; Klein et al., 2014; Open Science Collaboration, 2015; Prinz, Schlange & Asadullah, 2011). For example, an initiative by Bayer Healthcare to replicate 67 pre-clinical studies led to a reproducibility rate of 20-25% (Prinz et al., 2011), and researchers at Amgen were only able to replicate 6 of 53 influential cancer biology studies (Begley & Ellis, 2012). More recently, across a number of crowdsourced replication initiatives in social psychology, the majority of independent replications failed to find the significant results obtained in the original report (Ebersole et al., 2015; Klein et al., 2014; Open Science Collaboration, 2015).

The process of replicating published work can be adversarial (Bohannon, 2014; Gilbert, 2014; Kahneman, 2014; Schnall 2014a/b/c; but see Matzke et al., 2015), with concerns raised that some replicators select findings from research areas in which they lack expertise and about which they hold skeptical priors (Lieberman, 2014; Wilson, 2014). Some replications may have been conducted with insufficient pretesting and tailoring of study materials to the new subject population, or involve sensitive manipulations and measures that even experienced investigators find difficult to properly carry out (Mitchell, 2014; Schwarz & Strack, 2014; Stroebe & Strack, 2014). On the other side of the replication process, motivated reasoning (Ditto & Lopez, 1992; Kunda, 1990; Lord, Ross, & Lepper, 1979; Molden & Higgins, 2012) and post-hoc dissonance

(Festinger, 1957) may lead the original authors to dismiss evidence they would have accepted as disconfirmatory had it been available in the early stages of theoretical development (Mahoney, 1977; Nickerson, 1999; Schaller, in press; Tversky & Kahneman, 1971).

We introduce a complementary and more collaborative approach, in which findings are replicated in qualified independent laboratories before (rather than after) they are published (see also Schooler, 2014). We illustrate the Pre-Publication Independent Replication (PPIR) approach through a crowdsourced project in which 25 laboratories from around the world conducted replications of all 10 moral judgment effects which the last author and his collaborators had "in the pipeline" as of August 2014.

Each of the 10 original studies from Uhlmann and collaborators obtained support for at least one major theoretical prediction. The studies used simple designs that called for ANOVA followed up by t-tests of simple effects for the experiments, and correlational analyses for the nonexperimental studies. Importantly, for all original studies, all conditions and outcomes related to the theoretical hypotheses were included in the analyses, and no participants were excluded. Furthermore, with the exception of two studies that were run before 2010, results were only analyzed after data collection had been terminated. In these two older studies (*intuitive economics effect* and *belief-act inconsistency effect*), data were analyzed twice, once approximately halfway through data collection and then again after the termination of data collection. Thus, for most of the studies, a lack of replicability cannot be easily accounted for by exploitation of researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011) in the original analyses. For the purposes of transparency, the data and materials from both the original and replication studies are posted on the Open Science Framework (https://osf.io/q25xa/),

creating a publicly available resource researchers can use to better understand reproducibility and non-reproducibility in science (Kitchin, 2014; Nosek & Bar-Anan, 2012; Nosek, Spies, & Motyl, 2012).

In addition, as replicator labs were explicitly selected for their expertise and access to subject populations that were theoretically expected to exhibit the original effects, several of the most commonly given counter-explanations for failures to replicate are addressed by the present project. Under these conditions, a failure to replicate is more clearly attributable to the original study overestimating the relevant effect size— either due to the "winner's curse" suffered by underpowered studies that achieve significant results largely by chance (Button et al., 2013; Ioannidis, 2005; Ioannidis & Trikalinos, 2007; Schooler, 2011) or due to unanticipated differences between participant populations, which would suggest that the phenomenon is less general than initially hypothesized (Schaller, in press). Because one replication team was located at the institution where four of the original studies were run (Northwestern University), and replication laboratories are spread across six countries (United States, Canada, the Netherlands, France, Germany and China) it is possible to empirically assess the extent to which declines in effect sizes between original and replication studies (Schooler, 2011) are due to unexpected yet potentially meaningful cultural and population differences (Henrich, Heine, & Norenzayan, 2010). For all of these reasons, this crowdsourced PPIR initiative features replications arguably higher in informational value (Nelson, 2005; Nelson, McKenzie, Cottrell, & Sejnowski, 2010) than in prior work.

Method

Selection of Original Findings and Replication Laboratories

Ten moral judgment studies were selected for replication. We defined our sample of studies as all unpublished moral judgment effects the last author and his collaborators had "in the pipeline" as of August 2014. These moral judgment studies were ideal for a crowdsourced replication project because the study scenarios and dependent measures were straightforward to administer, and did not involve sensitive manipulations of participants' mindset or mood. They also examined basic aspects of moral psychology such as character attributions, interpersonal trust, and fairness that were not expected to vary dramatically between the available samples of research participants. All 10 original studies found support for at least one key theoretical prediction. The crowdsourced PPIR project assessed whether support for the same prediction was obtained by other research groups.

Unlike any previous large-scale replication project, all original findings targeted for replication in the Pipeline Project were unpublished rather than published. In addition, all findings were volunteered by the original authors, rather than selected by replicators from a set of prestigious journals in a given year (e.g., Open Science Collaboration, 2015) or nominated on a public website (e.g., Ebersole et al., 2015). In a further departure from previous replication initiatives, participation in the Pipeline Project was by invitation-only, via individual recruitment e-mails. This ensured that participating laboratories had both adequate expertise and access to a subject population in which the original finding was theoretically expected to replicate using the original materials (i.e., without any need for further pre-testing or revisions to the manipulations, scenarios, or dependent variables; Schwarz & Strack, 2014; Stroebe & Strack, 2014). Thus, the

PPIR project did not repeat the original studies in new locations without regard for context. Indeed, replication labs and locations were selected *a priori* by the original authors as appropriate to test the effect of interest.

Data Collection Process

Each independent laboratory conducted a direct replication of between three and ten of the targeted studies ($M_{studies}$ = 5.64, *SD* = 1.24), using the materials developed by the original researchers. To reduce participant fatigue, studies were typically administered using a computerized script in one of three packets, each containing three to four studies, with study-order counterbalanced between-subjects. There were four noteworthy exceptions to this approach. First, the Northwestern University replications were conducted using paper-pencil questionnaires, and participants were randomly assigned to complete a packet including either one longer study or three shorter studies in a fixed rather than counterbalanced order. Second, the Yale University replications omitted one study from one packet out of investigator concerns that the participant population might find the moral judgment scenario offensive. Third, the INSEAD Paris lab data collections included a translation error in one study that required it to be re-run separately from the others. Fourth and finally, the HEC Paris replication pack included six studies in a fixed order. Tables S1a-S1f in Supplement 1 summarize the replication locations, specific studies replicated, sample sizes, mode of study administration (online vs. laboratory), and type of subject population (general population, MBA students, graduate students, or undergraduate students) for each replication packet.

Replication packets were translated from English into the local language (e.g., Chinese, French) with the exception of the HEC Paris and Groningen data collections, where the materials

were in English, as participants were enrolled in classes delivered in English. All translations were checked for accuracy by at least two native language speakers. The complete Qualtrics links for the replication packets are available at https://osf.io/q25xa/.

We used a similar data collection process to the Many Labs initiatives (Klein et al., 2014; Ebersole et al., 2015). The studies were programmed and carefully error-checked by the project coordinators, who then distributed individual links to each replication team. Each participant's data was sent to a centralized Qualtrics account as they completed the study. After the data were collected, the files were compiled by the project's first and second author and uploaded to the Open Science Framework. We could not have replication teams upload their data directly to the OSF because it had to be carefully anonymized first. The Pipeline Project website on the OSF includes three master files, one for each pipeline packet, with the data from all the replication studies together. The data for the original studies is likewise posted, also in an anonymized format.

Replication teams were asked to collect 100 participants or more for at least one packet of replication studies. Although some of the individual replication studies had less than 80% power to detect an effect of the same size as the original study, aggregating samples across locations of our studies fulfilled Simonsohn's (2015) suggestion that replications should have at least 2.5 times as many participants as the original study. The largest-N original study collected data for 265 subjects; the aggregated replication samples for each finding range from 1542 participants to 3647 participants. Thus, the crowdsourced project allowed for high-powered tests of the original hypotheses and more accurate effect size estimates than the original data collections.

Specific Findings Targeted for Replication

Although the principal goal of the present article is to introduce the PPIR method and demonstrate its feasibility and effectiveness, each of the ten original effects targeted for replication are of theoretical interest in-and-of themselves. Detailed write-ups of the methods and results for each original study are provided in Supplement 2, and the complete replication materials are included in Supplement 3.

The bulk of the studies test core predictions of the person-centered account of moral judgment (Landy & Uhlmann, 2015; Pizarro & Tannenbaum, 2011; Pizarro, Tannenbaum, & Uhlmann, 2012; Uhlmann, Pizarro, & Diermeier, 2015). Two further studies explore the effects of moral concerns on perceptions of economic processes, with an eye toward better understanding public opposition to aspects of the free market system (Blendon et al., 1997; Caplan, 2001, 2002). A final pair of studies examine the implications of the psychology of moral judgment for corporate reputation management (Diermeier, 2011).

Person-centered morality. The person-centered account of moral judgment posits that moral evaluations are frequently driven by informational value regarding personal character rather than the harmfulness and blameworthiness of acts. As a result, less harmful acts can elicit more negative moral judgments, as long as they are more informative about personal character. Further, *act-person dissociations* can emerge, in which acts that are rated as less blameworthy than other acts nonetheless send clearer signals of poor character (Tannenbaum, Uhlmann, & Diermeier, 2011; Uhlmann, Zhu, & Tannenbaum, 2013). More broadly, the person-centered approach is consistent with research showing that internal factors, such as intentions, can be weighed more heavily in moral judgments than objective external consequences. The first six

studies targeted for replication in the Pipeline Project test ideas at the heart of the theory, and further represent the body of unpublished work from this research program. Large-sample failures to replicate many or most of these findings across 25 universities would at a minimum severely undermine the theory, and perhaps even invalidate it entirely. Brief descriptions of each of the person-centered morality studies are provided below.

Study 1: Bigot-Misanthrope Effect. Participants judge a manager who selectively mistreats racial minorities as a more blameworthy person than a manager who mistreats all of his employees. This supports the hypothesis that the informational value regarding character provided by patterns of behavior plays a more important role in moral judgments than aggregating harmful vs. helpful acts (Pizarro & Tannenbaum, 2011; Shweder & Haidt, 1993; Yuill, Perner, Pearson, Peerbhoy, & Ende, 1996).

Study 2: Cold-Hearted Prosociality Effect. A medical researcher who does experiments on animals is seen as engaging in more morally praiseworthy acts than a pet groomer, but also as a worse person. This effect emerges even in joint evaluation (Hsee, Loewenstein, Blount, & Bazerman, 1999), with the two targets evaluated at the same time. Such *act-person dissociations* demonstrate that moral evaluations of acts and the agents who carry them out can diverge in systematic and predictable ways. They represent the most unique prediction of, and therefore strongest evidence for, the person centered approach to moral judgment.

Study 3: Bad Tipper Effect. A person who leaves the full tip entirely in pennies is judged more negatively than a person who leaves less money in bills, and tipping in pennies is seen as higher in informational value regarding character. Like the bigot-misanthrope effect described above, this provides rare direct evidence of the role of perceived informational value regarding

character in moral judgments. Moral reactions often track perceived character deficits rather than harmful consequences (Pizarro & Tannenbaum, 2011; Yuill et al., 1996).

Study 4: Belief-Act Inconsistency Effect. An animal rights activist who is caught hunting is seen as an untrustworthy and bad person, even by participants who think hunting is morally acceptable. This reflects person centered morality: an act seen as morally permissible in-and-of-itself nonetheless provokes moral opprobrium due to its inconsistency with the agent's stated beliefs (Monin & Merritt, 2012; Valdesolo & DeSteno, 2007).

Study 5: Moral Inversion Effect. A company that contributes to charity but then spends even more money promoting the contribution in advertisements not only nullifies its generous deed, but is perceived even more negatively than a company that makes no donation at all. Thus, even an objectively helpful act can provoke moral condemnation, so long as it suggests negative underlying traits such as insincerity (Jordan, Diermeier, & Galinsky, 2012).

Study 6: Moral Cliff Effect. A company that airbrushes the model in their skin cream advertisement to make her skin look perfect is seen as more dishonest, ill-intentioned, and deserving of punishment than a company that hires a model whose skin already looks perfect. This theoretically reflects inferences about underlying intentions and traits (Pizarro & Tannenbaum, 2011; Yuill et al., 1996). In the two cases consumers have been equally misled by a perfect-looking model, but in the airbrushing case the deception seems more deliberate and explicitly dishonest.

Morality and markets. Studies 7 and 8 examined the role of moral considerations in lay perceptions of capitalism and businesspeople, in an effort to better understand discrepancies between the policy prescriptions of economists and everyday people (Blendon et al., 1997;

Caplan, 2001, 2002). Despite the material wealth created by free markets, moral intuitions lead to deep psychological unease with the inequality and incentive structures of capitalism. Understanding such intuitions is critical to bridging the gap between lay and scientific understandings of economic processes.

Study 7: Intuitive Economics Effect. Economic variables that are widely regarded as unfair are perceived as especially bad for the economy. Such a correlation raises the possibility that moral concerns about fairness irrationally influence perceptions of economic processes. In other words, aspects of free markets that seem unfair on moral grounds (e.g., replacing hard-working factory workers with automated machinery that can do the job more cheaply) may be subject to distorted perceptions of their objective economic effects (a moral coherence effect; Clark, Chen, & Ditto, in press; Liu & Ditto, 2013).

Study 8: Burn-in-Hell Effect. Participants perceive corporate executives as more likely to burn in hell than members of social categories defined by antisocial behavior, such as vandals. This of course reflects incredibly negative assumptions about senior business leaders. “Vandals” is a social category defined by bad behavior; “corporate executive” is simply an organizational role. However, the *assumed* behaviors of a corporate executive appear negative enough to warrant moral censure.

Reputation management. The final two studies examined how prior assumptions and beliefs can shape moral judgments of organizations faced with a reputational crisis. Corporate leaders may frequently fail to anticipate the negative assumptions everyday people have about businesses, or the types of moral standards that are applied to different types of organizations.

These issues hold important applied implications, given the often devastating economic consequences of psychologically misinformed reputation management (Diermeier, 2011).

Study 9: Presumption of Guilt Effect. For a company, failing to respond to accusations of misconduct leads to similar judgments as being investigated and found guilty. If companies accused of wrongdoing are simply assumed to be guilty until proven otherwise, this means that aggressive reputation management during a corporate crisis is imperative. Inaction or "no comment" responses to public accusations may be in effect an admission of guilt (Fehr & Gelfand, 2010; Pace, Fediuk, & Botero, 2010).

Study 10: Higher Standard Effect. It is perceived as acceptable for a private company to give small (but not large) perks to its top executive. But for the leader of a charitable organization, even a small perk is seen as moral transgression. Thus, under some conditions a praiseworthy reputation and laudable goals can actually hurt an organization, by leading it to be held to a higher moral standard.

The original data collections found empirical support for each of the ten effects described above. These studies possessed many of the same strengths and weaknesses found in the typical published research in social psychology journals. On the positive side, the original studies featured hypotheses strongly grounded in prior theory, and research designs that allowed for excellent experimental control (and for the experimental designs, causal inferences). On the negative side, the original studies had only modest sample sizes and relied on participants from one subpopulation of a single country. The Pipeline Project assessed how many of these findings would emerge as robust in large-sample crowdsourced replications at universities across the world.

Pre-Registered Analysis Plan

The pre-registration documents for our analyses are posted on the Open Science Framework (https://osf.io/uivsj/), and are also included in Supplement 4. The HEC Paris replications were conducted prior to the pre-registration of the analyses for the project; all of the other replications can be considered pre-registered (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). None of the original studies were pre-registered.

Pre-registration is a new method in social psychology (Nosek & Lakens, 2014; Wagenmakers et al., 2012), and there is currently no fixed standard of practice. It is generally acknowledged that, when feasible, registering a plan for how the data will be analyzed in advance is a good research practice that increases confidence in the reported findings. As with prior crowdsourced replication projects (Ebersole et al., 2015; Klein et al., 2014; Open Science Collaboration, 2015), the present report focuses on the primary effect of interest from each original study. Our pre-registration documents therefore list 1) the replication criteria that will be applied to the original findings and 2) the specific comparisons used to calculate the replication effect sizes (i.e., for each study, which condition will be compared to which condition and what the key dependent measure is). This method provides few researcher degrees of freedom (Simmons et al., 2011) to exaggerate the replicability of the original findings.

Results

Data Exclusion Policy

None of the original studies targeted for replication dropped observations, and to be consistent none of the replication studies did either. It is true in principle that excluding data is sometimes justified and can lead to more accurate inferences (Berinsky, Margolis, & Sances, in

press; Curran, in press). However, in practice this is often done in the post-hoc manner and exploits researcher degrees of freedom to produce false positive findings (Simmons et al., 2011). Excluding observations is most justified in the case of research highly subject to noisy data, such as research on reaction times for instance (Greenwald, Nosek, & Banaji, 2003). The present studies almost exclusively used simple moral judgment scenarios and self-report Likert-type scale responses for dependent measures. Our policy was therefore to not remove any participants from the analyses of the replication results, and dropping observations was not part of the pre-registered analysis plan for the project. The data from the Pipeline Project is publicly posted on the Open Science Framework website, and anyone interested can re-analyze the data using participant exclusions if they wish to do so.

Original and Replication Effect Sizes

The original effect sizes, meta-analyzed replication effect sizes, and effect sizes for each individual replication are shown in Figure 1. To obtain the displayed effect sizes, a standardized mean difference (*d*, Cohen, 1988) was computed for each study and each sample taking the difference between the means of the two sets of scores and dividing it by the sample standard deviation (for uncorrelated means) or by the standard deviation of the difference scores (for dependent means). Effect sizes for each study were then combined according to a random effects model, weighting every single effect for the inverse of its total variance (i.e. the sum of the between- and within-study variances) (Cochran & Carroll, 1953; Lipsey & Wilson, 2001). The variances of the combined effects were computed as the reciprocal of the sum of the weights and the standard error of the combined effect sizes as the squared root of the variance, and 95% confidence intervals were computed by adding to and subtracting from the combined effect 1.96

standard errors. Formulas for the calculation of meta-analytic means, variances, SEs, and CIs were taken from Borenstein, Hedges, Higgins and Rothstein (2011) and from Cooper and Hedges (1994).

In Figure 1 and in the present report more generally our focus is on the replicability of the single most theoretically important result from each original study (Ebersole et al., 2015; Klein et al., 2014; Open Science Collaboration, 2015). However, Supplement 2 more comprehensively repeats the analyses from each original study, with the replication results appearing in brackets in red font immediately after the same statistical test from the original investigation.

Applying Frequentist Replication Criteria

There is currently no single, fixed standard to evaluate replication results, and as outlined in our pre-registration plan we therefore applied five primary criteria to determine whether the replications successfully reproduced the original findings (Brandt et al., 2014; Simonsohn, 2015). These included whether 1) the original and replication effects were in the same direction, 2) the replication effect was statistically significant after meta-analyzing across all replications, 3) meta-analyzing the original and replication effects resulted in a significant effect, 4) the original effect was inside the confidence interval of the meta-analyzed replication effect size, and 5) the replication effect size was large enough to have been reliably detected in the original study ("small telescopes" criterion; Simonsohn, 2015). Table 1 evaluates the replication results for each study along the five dimensions, as well as providing a more holistic evaluation in the final column.

The small telescopes criterion perhaps requires some further explanation. A large-N replication can yield an effect size significantly different from zero that still fails the small

telescopes criterion, in that the "true" (i.e., replication) effect is too small to have been reliably detected in the original study. This suggests that although the results of the original study could just have been due to statistical noise, the original authors did correctly predict the direction of the true effect from their theoretical reasoning (see also Hales, in press). Figure S5 in Supplement 5 summarizes the small telescopes results, including each original effect size, the corresponding aggregated replication effect size, and d33% line indicating the smallest effect size that would be reasonably detectable with the original study design.

Interpreting the replication results is slightly more complicated for the two original findings which involved null effects (*presumption of guilt effect* and *higher standard effect*). The original presumption of guilt study found that failing to respond to accusations of wrongdoing is perceived equally negatively as being investigated and found guilty. The original higher standard study predicted and found that receiving a small perk (as opposed to purely monetary compensation) negatively affected the reputation of the head of a charity (significant effect), but not a corporate executive (null effect), consistent with the idea that charitable organizations are held to a higher standard than for-profit companies. For these two studies, a failure to replicate would involve a significant effect emerging where there had not been one before, or a replication effect size significantly *larger* than the original null effect. The more holistic evaluation of replicability in the final column of Table 1 takes these nuances into account.

Meta-analyzed replication effects for all studies were in the same direction as the original effect (see Table 1). In eight out of ten studies, effects that were significant or nonsignificant in the original study were likewise significant or nonsignificant in the crowdsourced replication (eight of eight effects that were significant in the original study were also significant in the

replication; neither of the two original findings that were null effects in the original study were also null effects in the replication). Including the original effect size in the meta-analysis did not change these results, due to the much larger total sample of the crowdsourced replications. For four out of ten studies, the confidence interval of the meta-analyzed replication effect did not include the original effect; in one case, this occurred because the replication effect was smaller than the original effect. No study failed the small telescopes criterion.

Applying a Bayesian Replication Criterion

Replication results can also be evaluated using Bayesian methods (e.g., Verhagen & Wagenmakers, 2014; Wagenmakers, Verhagen, & Ly, in press). Here we focus on the Bayes factor, a continuous measure of evidence that quantifies the degree to which the observed data are predicted better by the alternative hypothesis than by the null hypothesis (e.g., Jeffreys, 1961; Kass & Raftery, 1995; Rouder et al., 2009; Wagenmakers, Grünwald, & Steyvers, 2006). The Bayes factor requires that the alternative hypothesis is able to make predictions, and this necessitates that its parameters are assigned specific values. In the Bayesian framework, the uncertainty associated with these values is described through a prior probability distribution. For instance, the default correlation test features an alternative hypothesis that stipulates all values of the population correlation coefficient to be equally likely a priori – this is instantiated by a uniform prior distribution that ranges from -1 to 1 (Jeffreys, 1961; Ly, Verhagen, & Wagenmakers, in press). Another example is the default t-test, which features an alternative hypothesis that assigns a fat-tailed prior distribution to effect size (i.e., a Cauchy distribution with scale r=1; for details see Jeffreys, 1961; Ly et al., in press; Rouder et al., 2009).

In a first analysis, we applied the default Bayes factor hypothesis tests to the original findings. The results are shown on the y-axis of Figure 2. For instance, the original experiment on the *moral inversion effect* yielded $BF_{10} = 7.526$, meaning that the original data are about 7.5 times more likely under the default alternative hypothesis than under the null hypothesis. For 9 out of 11 effects, $BF_{10} > 1$, indicating evidence in favor of the default alternative hypothesis. This evidence was particularly compelling for the following five effects: the *moral cliff effect*, the *cold hearted prosociality effect*, the *bigot-misanthrope effect*, the *intuitive economics effect*, and – albeit to a lesser degree– the *higher standards-charity effect*. The evidence was less conclusive for *the moral inversion effect*, the *bad tipper effect*, and the *burn in hell effect*; for the *belief-act inconsistency effect*, the evidence is almost perfectly ambiguous (i.e., $BF_{10} = 1.119$). In contrast, and as predicted by the theory, the Bayes factor indicates support in favor of the null hypothesis for the *presumption of guilt effect* (i.e., $BF_{01} = 5.604$; note the switch in subscripts: the data are about 5.6 times more likely under the null hypothesis than under the alternative hypothesis). Finally, for the *higher standards-company effect* the Bayes factor does signal support in favor of the null hypothesis – as predicted by the theory – but only by a narrow margin (i.e., $BF_{01} = 1.781$).

In a second analysis, we take advantage of the fact that Bayesian inference is, at its core, a theory of optimal learning. Specifically, in order to gauge replication success we calculate Bayes factors separately for each replication study; however, we now depart from the default specification of the alternative hypothesis and instead use a highly informed prior distribution, namely the posterior distribution from the original experiment. This informed alternative hypothesis captures the belief of a rational agent who has seen the original experiment and

believes the alternative hypothesis to be true. In other words, our replication Bayes factors contrast the predictive adequacy of two hypotheses: the standard null hypothesis that corresponds to the belief of a skeptic and an informed alternative hypothesis that corresponds to the idealized belief of a proponent (Verhagen & Wagenmakers, 2014; Wagenmakers et al., in press).

The replication Bayes factors are denoted by BF_{r0} and are displayed by the grey dots in Figure 2. Most informed alternative hypotheses received overwhelming support, as indicated by very high values of BF_{r0}. From a Bayesian perspective, the original effects therefore generally replicated with a few exceptions. First, the replication Bayes factors favor the informed alternative hypothesis for the *higher standards-company effect*, when the original was a null finding. Second, the evidence is neither conclusively for nor against the *presumption of guilt effect*, which was also originally a null finding. The data and output for the Bayesian assessments of the original and replication results are available on the Pipeline Project's OSF page: https://osf.io/q25xa/.

Moderator Analyses

Table 2 summarizes whether a number of sample characteristics and methodological variables significantly moderated the replication effect sizes for each original finding targeted for replication. Supplement 6 provides more details on our moderator analyses.

USA vs. non-USA sample. As noted earlier, no cultural differences for any of the original findings were hypothesized *a priori*. In fact replication laboratories were chosen for the PPIR initiative due to their access to subject populations in which the original effect was theoretically predicted to emerge. However, it is, of course, an empirical question whether the effects vary across cultures or not. Since all of the original studies were conducted in the United States, we

examined whether replication location (USA vs. non-USA) made a difference. As seen in Table 2, six out of ten original findings exhibited significantly larger effect sizes in replications in the United States, whereas the reverse was true for one original effect.

Results for one original finding, the *bad tipper effect*, were especially variable across cultures ($Q(16) = 165.39$, $p < .001$). The percentage of variability between studies that is due to heterogeneity rather than chance (random sampling) is $I^2 = 90.32\%$ in the bad tipper effect and 68.11% on average in all other effects studied. The *bad tipper effect* is the only study in which non-US samples account for approximately a half of the total true heterogeneity. The drop in I^2 that we observed when we removed non-USA samples from the other studies was 10.78% on average. The bad tipper effect replicated consistently in the United States ($d_{usa} = 0.74$, 95% CI [.62, .87]), but less so outside the USA ($d_{non\text{-}usa} = 0.30$, 95% CI [-.22, .82])[1]; the difference between these two effect sizes was significant, $F(1,3635) = 52.59$, $p = .01$. The bad tipper effect actually significantly *reversed* in the replication from the Netherlands ($d_{netherlands} = -0.46$, $p < .01$). Although post hoc, one potential explanation could be cultural differences in tipping norms (Azar, 2007). Regardless of the underlying reasons, these analyses provide initial evidence of cultural variability in the replication results.

Same vs. different location as original study. To our knowledge, the Pipeline Project is the only crowdsourced replication initiative to systematically re-run all of the targeted studies using the original subject population. For instance, four studies conducted between 2007 and 2010 using Northwestern undergraduates as participants were replicated in 2015 as part of the project, again using Northwestern undergraduates. We reran our analyses of key effects and included study location as a moderating variable (different location as original study = coded as

0; same location as original study = coded as 1). This allowed us to examine whether findings were more likely to replicate in the original population than in new populations. As seen in Table 2, four effects were significantly larger in the original location, two effects were actually significantly larger in locations *other than* the original study site, and for five effects same versus different location was not a significant moderator.

Student sample vs. general population. The type of subject population was likewise examined. A general criticism of psychological research is its over-reliance on undergraduate student samples, arguably limiting the generalizability of research findings (Sears, 1986). As seen in Table 2, five effects were larger in the general population than in student samples, whereas the reverse was true for one effect.

Study order. As subject fatigue may create noise and reduce estimated effect sizes, we examined whether the order in which the study appeared made a difference. It seemed likely that studies would be more likely to replicate when administered earlier in the replication packet. However, order only significantly moderated replication effect sizes for one finding, which was unexpectedly larger when included *later* in the replication packet.

Holistic Assessment of Replication Results

Given the complexity of replication results and the plethora of criteria with which to evaluate them (see Brandt et al., 2014; Simonsohn, 2015), we close with a holistic assessment of the results of this first Pre-Publication Independent Replication (PPIR) initiative (see also the final column of Table 1).

Six out of ten of the original findings replicated quite robustly across laboratories: the *bigot-misanthrope effect, belief-act inconsistency effect, cold-hearted prosociality, moral cliff*

effect, burn in hell effect, and *intuitive economics effect*. For these original findings the aggregated replication effect size was 1) in the same direction as in the original study, 2) statistically significant after meta-analyzing across all replications, 3) significant after meta-analyzing across both the original and replication effect sizes, 4) not significantly different from the original effect size, and 5) large enough to have been reliably detected in the original study ("small telescopes" criterion; Simonsohn, 2015).

The *bad tipper effect* likewise replicated according to the above criteria, but with some evidence of moderation by national culture. According to the frequentist criteria of statistical significance, the effect replicated consistently in the United States, but not in international samples. This could be due to cultural differences in norms related to tipping. It is noteworthy however that in the Bayesian analyses, almost all replication Bayes factors favor the original hypothesis, suggesting there is a true effect.

The *moral inversion effect* is another interesting case. This effect was statistically significant in both the original and replication studies. However, the replication effect was smaller and with a confidence interval that did not include the original effect size, suggesting the original study overestimated the true effect. Yet despite this, the moral inversion effect passed the small telescopes criterion (Simonsohn, 2015): the aggregated replication effect was large enough to have been reliably detected in the original study. The original study therefore provided evidence for the hypothesis that was unlikely to be mere statistical noise. Overall, we consider the moral inversion effect supported by the crowdsourced replication initiative.

In contrast, two findings failed to consistently replicate the same pattern of results found in the original study (*higher standards* and *presumption of guilt*). We consider the original

theoretical predictions not supported by the large-scale PPIR project. These studies were re-run in qualified labs using subject populations predicted *a priori* by the original authors to exhibit the hypothesized effects, and failed to replicate in high-powered research designs that allowed for much more accurate effect size estimates than in the original studies. Notably, both of these original studies found null effects; the crowdsourced replications revealed true effect sizes that were both significantly different from zero and two to five times larger than in the original investigations. Replication Bayes factors for the higher standards effect suggest the original finding is simply not true, whereas results for the presumption of guilt hypothesis are ambiguous, suggesting the evidence is not compelling either way.

As noted earlier, the original effects examined in the Pipeline Project fell into three broad categories: person-centered morality, moral perceptions of market factors, and the psychology of corporate reputation. These three broad categories of effects received differential support from the results of the crowdsourced replication initiative. Robust support for predictions based on the person-centered account of moral judgment (Pizarro & Tannenbaum, 2011; Uhlmann et al., 2015) was obtained across six original findings. The replication results also supported predictions regarding moral coherence (Liu & Ditto, 2013; Clark et al., in press) in perceptions of economic variables, but did not find consistent support for two specific hypotheses concerning moral judgments of organizations. Thus, the PPIR process allowed categories of robust effects to be separated from categories of findings that were less robust. Although speculative, it may be the case that programmatic research supported by many conceptual replications is more likely to directly replicate than stand-alone findings (Schaller, in press).

Discussion

The present crowdsourced project introduces Pre-Publication Independent Replication (PPIR), a method for improving the reproducibility of scientific research in which findings are replicated in qualified independent laboratories before (rather than after) they are published. PPIRs are high in informational value (Nelson, 2005; Nelson et al., 2010) because replication labs are chosen by the original author based on their expertise and access to relevant subject populations. Thus, common alternative explanations for failures to replicate are eliminated, and the replication results are especially diagnostic of the validity of the original claims.

The Pre-Publication Independent Replication approach is a practical method to improve the rigor and replicability of the published literature that complements the currently prevalent approach of replicating findings after the original work has been already published (Ebersole et al., 2015; Klein et al., 2014; Open Science Collaboration, 2015). The PPIR method increases transparency and ensures that what appears in our journals has already been independently validated. PPIR avoids the cluttering of journal archives by original articles and replication reports that are separated in time, and even publication outlets, without any formal links to one another. It also avoids the adversarial interactions that sometimes occur between original authors and replicators (see Bohannon, 2014; Kahneman, 2014; Lieberman, 2014; Schnall 2014a/b/c; Wilson, 2014). In sum, the PPIR approach represents a powerful example of how to conduct a reliable and effective science that fosters capitalization rather than self-correction.

To illustrate the PPIR approach, 25 research groups conducted replications of all moral judgment effects which the last author and his collaborators had "in the pipeline" as of August 2014. Six of the ten original findings replicated robustly across laboratories. At the same time,

four of the original findings failed at least one important replication criterion (Brandt et al., 2014; Schaller, in press; Verhagen & Wagenmakers, 2014)— either because the effect only replicated significantly in the original culture (one study), because the replication effect was significantly smaller than in the original (one study), because the original finding consistently failed to replicate according to frequentist criteria (two studies), or because the replication Bayes factor disfavored the original finding (one study) or revealed mixed results (one study).

Moderate Replication Rates Should Be Expected

The overall replication rate in the Pipeline Project was higher than in the Reproducibility Project (Open Science Collaboration, 2015), in which 36% of 100 original studies were successfully replicated at the $p < .05$ level. (Notably, 47% of the original effect sizes fell within the 95% confidence interval of the replication effect size). However, the two crowdsourced replication projects are not directly comparable, since the Reproducibility Project featured single-shot replications from individual laboratories and the Pipeline Project pooled the results of multiple labs, leading to greater statistical power to detect small effects. As seen in Figure 1, the individual labs in the Pipeline Project often failed to replicate original findings that proved reliable when the results of multiple replication sites were combined. In addition, the moral judgment findings in the Pipeline Project required less expertise to carry out than some of the original findings featured in the Reproducibility Project, which likely also improved our replication rate.

A more relevant comparison is the Many Labs projects, which pioneered the use of multiple laboratories to replicate the same findings. Our replication rate of 60%-80% was comparable to the Many Labs 1 project, in which eleven out of thirteen effects replicated across

36 laboratories (Klein et al., 2014). However it was higher than in Many Labs 3, in which only 30% of ten original studies replicated across 20 laboratories (Ebersole et al., 2015). Importantly, both original studies from our Pipeline Project that failed to replicate (*higher standard* and *presumption of guilt*), as well as the studies which replicated with highly variable results across cultures (*bad tipper*) or obtained a smaller replication effect size than the original (*moral inversion*) featured open data and materials. They also featured no use of questionable research practices such as optional stopping, failing to report all dependent measures, or removal of subjects, conditions, and outliers (Simmons et al., 2011). This underscores the fact that studies will often fail to replicate for reasons having nothing to do with scientific fraud or questionable research practices.

It is important to remember that null hypothesis significance testing establishes a relatively low threshold of evidence. Thus, many effects that were supported in the original study *should not* find statistical support in replication studies (Stanley & Spence, 2014), especially if the original study was underpowered (Cumming, 2008; Simmons, Nelson, & Simonsohn, 2013) and the replication relied on participants from a different culture or demographic group who may have interpreted the materials differently (Fabrigar & Wegener, in press; Stroebe, in press). In the present crowdsourced PPIR investigation, the replication rate was imperfect despite original studies that were conducted transparently and re-examined by qualified replicators. It may be expecting too much for an effect obtained in one laboratory and subject population to automatically replicate in any other laboratory and subject population.

Increasing sample sizes dramatically (for instance, to 200 subjects per cell) reduces both Type 1 and Type 2 errors by increasing statistical power, but may not be logistically or

economically feasible for many research laboratories (Schaller, in press). Researchers could instead run initial studies with moderate sample sizes (for instance, 50 subjects per condition in the case of an experimental study; Simmons et al., 2013), conduct similarly powered self-replications, and then explore the generality and boundary conditions of the effect in large-sample crowdsourced PPIR projects. This is a variation of the *Explore Small, Confirm Big* strategy proposed by Sakaluk (in press).

It may also be useful to consider a Bayesian framework for evaluating replication results. Instead of forcing a yes-no decision based on statistical significance, in which nonsignificant results are interpreted as failures to replicate, a replication Bayes factor allows us to assess the degree to which the evidence supports the original effect *or* the null hypothesis (Verhagen & Wagenmakers, 2014).

Limitations and Challenges of Pre-Publication Independent Replication

It is important to also consider disadvantages of seeking to independently replicate findings prior to publication. Although more collegial and collaborative than replications of published findings (Bohannon, 2014), PPIRs do not speak to the reproducibility of classic and widely influential findings in the field, as is the case for instance with the Many Labs investigations (Ebersole et al., 2015; Klein et al., 2014). Rather, PPIRs can help scholars ensure the validity and reproducibility of their emerging research streams. The benefit of PPIR is perhaps clearest in the case of "hot" new findings celebrated at conferences and likely headed toward publication in high-profile journals and widespread media coverage. In these cases, there is enormous benefit to ensuring the reliability of the work before it becomes common knowledge among the general public. Correcting unreliable, but widely disseminated, findings post-

publication (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012) is much more difficult than systematically replicating new findings in independent laboratories before they appear in print (Schooler, 2014).

Critics have suggested that effect sizes in replications of already published work may be biased downward by lack of replicator expertise, use of replication samples where the original effect would not be theoretically expected to emerge (at least without further pre-testing), and confirmation bias on the part of skeptical replicators (Lieberman, 2014; Schnall, 2014a/b/c; Schwarz & Strack, 2014; Stroebe & Strack, 2014; Wilson, 2014). PPIR explicitly recruits expert partner laboratories with relevant participant populations and is less subject to these concerns. However, PPIRs could potentially suffer from the reverse problem, in other words estimated effect sizes that are *upwardly* biased. In a PPIR, replicators likely begin with the hope of confirming the original findings, especially if they are co-authors on the same research report or are part of the same social network as the original authors. But at the same time, the replication analyses are pre-registered, which dramatically reduces researcher degrees of freedom to drop dependent measures, strategically remove outliers, selectively report studies, and so forth. The replication datasets are further publicly posted on the internet. It is difficult to see how a replicator could artificially bias the result in favor of the original finding without committing outright fraud. The incentive to confirm the original finding in the PPIR may simply lead replicators to treat the study with the same care and professionalism that they would their own original work.

The respective strengths and weaknesses of PPIRs and replications of already published work can effectively balance one another. Initiatives such as Many Labs and the Reproducibility

Project speak to the reliability of already well-known and influential research; PPIRs provide a check against findings becoming too well-known and influential prematurely, before they are established as reliable.

Both existing replication methods (Klein et al., 2014) and PPIRs are best suited to simple studies that do not require very high levels of expertise to carry out, such as those in the present Pipeline Project. Many original findings suffer from a limited pool of qualified experts able and willing to replicate them. In addition, studies that involve sensitive manipulations can fail even in the hands of experts. For such studies, the informational value of null effects will generally be lower than positive effects, since null effects could be either due to an incorrect hypothesis or some aspect of the study not being executed correctly (Mitchell, 2014). Thus, although replication failures in the Pipeline Project have high informational value, not all future PPIRs will be so diagnostic of scientific truth.

Finally, it would be logistically challenging to implement pre-publication independent replication as a routine method for every paper in a researcher's pipeline. The number of new original findings emerging will invariably outweigh the PPIR opportunities. In practice, some lower-profile findings will face difficulties attracting replicator laboratories to carry out PPIRs. Researchers may pursue PPIRs for projects they consider particularly important and likely to be influential, or that have policy implications. We discuss potential ways to incentivize and integrate PPIRs into the current publication system below.

Implementing and Incentivizing PPIRs

Rather than advocate for mandatory independent replication prior to publication, we suggest that the improved credibility of findings that are independently replicated will constitute

an increasingly important quality signal in the coming years. As a field we can establish a premium for research that has been independently replicated prior to publication through favorable reviews and editorial decisions. Replicators can either be acknowledged formally as authors (with their role in the project made explicit in the author contribution statement) or a separate replication report can be submitted and published alongside the original paper. Research groups can also engage in "study swaps" in which they replicate each other's ongoing work.

Organizing independent replications across partner universities can be an arduous and time-consuming endeavor. Researchers with limited resources need a way to independently confirm the reliability of their work. To facilitate more widespread implementation of PPIRs, we plan to create a website where original authors can nominate their findings for PPIRs and post their study materials. Graduate methods classes all over the world will then adopt these for PPIR projects and the results will be published as the Pipeline Project 2 with the original researchers and replicators as co-authors. Additional information obtained from the replications (such as more precise measures of effect size, narrower confidence intervals, etc.) can then be incorporated into the final publications by the original authors with the replicators thanked in the acknowledgments. Obviously this approach is best suited to simple studies that require little expertise, such that a first-year graduate student can easily run the replications.

For original studies requiring high expertise and/or specialized equipment, one can envision a large online pool of interested laboratories, with expertise and resources publicly listed. The logic is similar to that of a temporary internet labor market, in which employers and workers in different parts of the world post profiles and find suitable matches through a bidding process. A similar “collaborator commons” for open science projects could be used to match

original laboratories seeking to have their work replicated with qualified experts.[2] Leveraging such an approach, even studies that require a great deal of expertise could be replicated independently prior to publication, so long as a suitable partner lab elsewhere in the world can be identified. The Many Lab website for online collaborations recently introduced by the Open Science Center (Ebersole, Klein, & Atherton, 2014) already provides the beginnings of such a system.

An online marketplace in which researchers offer up particular findings for replication can also help determine the interest and importance of the finding. Few will volunteer to help replicate an effect that is not interesting or novel. Thus a marketplace approach can not only help select out effects that are not reliable before publication, but also those that are less likely to capture broad interest from other researchers who study the same topic.

A challenge for pre-publication independent replication is credit and authorship. It is standard practice on crowdsourced replication projects to include replicators as co-authors (e.g., Alogna et al., 2014; Ebersole et al., 2015; Klein et al., 2014; Lai et al., 2014); we know of no exception to this principle. As with any large-scale collaborative project, author contributions are typically more limited than a traditional research publication, but this is proportional to the credit received— 54th author will gain no one an academic position or tenure promotion. Yet many colleagues still choose to take part, and large crowdsourced projects with long author strings have become increasingly common in recent years. This "many authors" approach is critical to the viability of crowdsourced research as a means to improve the rigor and replicability of our science. However, an extended author string can make it difficult to distinguish the relative

contributions of different project members. Detailed author contribution statements are critical to clarifying each person's respective project roles.

Integrating PPIRs and Cross-Cultural Research

Cross-cultural research bears important similarities with PPIRs, in that original studies are repeated in new populations by partner laboratories. Most research investigations unfortunately do not include cross-cultural comparisons (Henrich et al., 2010), leaving it an open question whether the observed phenomenon is similar or different outside the culture in which the research was originally done. It is worth considering how PPIRs and cross-cultural research can be better integrated to establish either the generalizability or cultural boundaries of a phenomenon.

Based on the theorizing underlying the ten effects selected for the Pipeline project, the original findings should have replicated consistently across laboratories. No cultural differences were hypothesized beforehand, yet such differences did emerge. For instance, a number of effects were significantly smaller outside of the original culture of the United States. We suggest that researchers conducting PPIRs include any anticipated moderation by replication location in their pre-registration plan (Wagenmakers et al., 2012). This allows for tests of *a priori* predictions regarding the cultural variability of a phenomenon. In cases such as ours in which the results point to unanticipated cultural differences, we suggest the investigators follow up with further confirmatory replications.

More ambitiously, large globally distributed PPIR initiatives could adopt a *replication chain* approach to probe the generalizability vs. cultural boundaries of an effect as efficiently as possible (see Hüffmeier, Mazei, & Schultze, in press, and Kahneman, 2012, for complementary

perspectives on how sequences of replications can inform theory and practice). In a replication chain, each original effect is first replicated in partner laboratories with subject populations as similar as possible to the original one. For instance, if a researcher at the University of Washington believes she has identified a reliable effect among UW undergraduates, the effect is first replicated at the University of Oregon and then other institutions in the United States. Only findings that replicate reliably within their original culture would then qualify for international replications at partner institutions in China, France, Singapore, and so on. There is little sense in expending limited resources on effects that do not consistently replicate even in similar subject populations.

Once an effect is established as reliable in its culture of origin, claims of cross-cultural universality can be put to a rigorous test by deliberately selecting the available replication sample *as different as possible* from the original subject population (Norenzayan & Heine, 2005). For instance, if a researcher at Princeton predicts that she has identified a universal judgmental bias, laboratories in non-Western cultures with access to less educated subject populations should be engaged for the PPIRs. A successful field replication (Maner, in press) among fishing net weavers in a rural village in China provides critical evidence of universality; a successful replication at Harvard adds very little evidence in support of this particular claim.

Alternatively, crowdsourced PPIR initiatives could coordinate laboratories to systematically test cultural moderators hypothesized *a priori* by the original authors (Henrich et al., 2010). Once the finding is established as reliable in its original culture, the authors select a specific culture or cultures for the replications in which the effect is expected to be absent or reverse on *a priori* grounds (e.g., Eastern vs. Western cultures; Nisbett, Peng, Choi, &

Norenzayan, 2001). Systematic tests across multiple universities in each culture then provide a safeguard against the nonrepresentative samples at each institution, which could confound comparisons if just one student population from each culture is used.

Conclusion

Pre-Publication Independent Replication is a collaborative approach to improving research quality in which original authors nominate their own unpublished findings and select expert laboratories to replicate their work. The aim is a replication process free of acrimony and final results high in scientific truth value. We illustrate the PPIR approach through crowdsourced replications of the last author's research pipeline, revealing the mix of robust, qualified, and culturally-variable effects that are to be expected when original studies and replications are conducted transparently. Integrating pre-publication independent replication into our research streams holds enormous potential for building connections between colleagues and increasing the robustness and reliability of scientific knowledge, whether in psychology or in other disciplines.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., Bornstein, B. H., Bouwmeester, S., Brandimonte, M. A., Brown, C., Buswell, K., Carlson, C., Carlson, M., Chu, S., Cislak, A., Colarusso, M., Colloff, M. F., Dellapaolera, K. S., Delvenne, J.-F., Di Domenico, A., Drummond, A., Echterhoff, G., Edlund, J. E., Eggleston, C. M., Fairfield, B., Franco, G., Gabbert, F., Gamblin, B. W., Garry, M., Gentry, R., Gilbert, E. A., Greenberg, D. L., Halberstadt, J., Hall, L., Hancock, P. J. B., Hirsch, D., Holt, G., Jackson, J. C., Jong, J., Kehn, A., Koch, C., Kopietz, R., Körner, U., Kunar, M. A., Lai, C. K., Langton, S. R. H., Leite, F. P., Mammarella, N., Marsh, J. E., McConnaughy, K. A., McCoy, S., McIntyre, A. H., Meissner, C. A., Michael, R. B., Mitchell, A. A., Mugayar-Baldocchi, M., Musselman, R., Ng, C., Nichols, A. L., Nunez, N. L., Palmer, M. A., Pappagianopoulos, J. E., Petro, M. S., Poirier, C. R., Portch, E., Rainsford, M., Rancourt, A., Romig, C., Rubínová, E., Sanson, M., Satchell, L., Sauer, J. D., Schweitzer, K., Shaheed, J., Skelton, F., Sullivan, G. A., Susa, K. J., Swanner, J. K., Thompson, W. B., Todaro, R., Ulatowska, J., Valentine, T., Verkoeijen, P. P. J. L., Vranka, M., Wade, K. A., Was, C. A., Weatherford, D., Wiseman, K., Zaksaite, T., Zuj, D. V., & Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556-578.

Azar, O. H. (2007). The social norm of tipping: A review. *Journal of Applied Social Psychology*, *37*(2), 380–402.

Begley, C.G., & Ellis, L.M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483,* 531–533.

Berinsky, A., Margolis, M.F., & Sances, M.W. (in press). Can we turn shirkers into workers? *Journal of Experimental Social Psychology.*

Blendon, R.J., Benson, J.M., Brodie, M., Morin, R., Altman, D.E., Gitterman, G., Brossard, M., & James, M. (1997). Bridging the gap between the public's and economists' views of the economy. *Journal of Economic Perspectives, 11,* 105-118.

Bohannon, J. (2014). Replication effort provokes praise—and 'bullying' charges. *Science, 344*, 788-789.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis.* John Wiley & Sons.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J., & van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217-224.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 1-12.

Caplan, B. (2001). What makes people think like economists? Evidence on economic cognition from the 'Survey of Americans and Economists on the Economy'. *Journal of Law and Economics, 43*, 395-426.

Caplan, B. (2002). Systematically biased beliefs about economics: Robust evidence of judgmental anomalies from the 'Survey of Americans and Economists on the Economy'. *Economic Journal, 112*, 433-458.

Clark, C. J., Chen, E. E., & Ditto, P. H. (in press). Moral coherence processes: Constructing

culpability and consequences. *Current Opinion in Psychology*.

Cochran, W. G., & Carroll, S. P. (1953). A sampling investigation of the efficiency of weighting inversely as the estimated variance. *Biometrics, 9(4),* 447-459.

Cohen, J. (1988), *Statistical power analysis for the behavioral sciences* (2nd ed.), New Jersey: Lawrence Erlbaum Associates.

Cooper, H., & Hedges, L. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286-300.

Curran, P.G. (in press). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*.

Davis-Stober, C. P., & Dana, J. (2014). Comparing the accuracy of experimental estimates to guessing: A new perspective on replication and the “crisis of confidence" in psychology. *Behavior Research Methods, 46,* 1-14.

Diermeier, D. (2011). *Reputation rules: Strategies for building your company’s most valuable asset*. McGraw-Hill.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: The use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*, 568–584.

Ebersole, C.R., Atherton, O.E., Belanger, A.L., Skulborstad, H.M. et al., & Nosek, B.A. (in

press). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*.

Ebersole, C.R., Klein, R.A., & Atherton, O.E. (2014). The Many Lab. https://osf.io/89vqh/https://osf.io/89vqh/

Fabrigar, L.R., & Wegener, D.T. (in press). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology.*

Fehr, R., & Gelfand, M. J. (2010). When apologies work: How matching apology components to victims' self-construals facilitates forgiveness. *Organizational Behavior and Human Decision Processes, 113*(1), 37–50.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford University Press.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (Sign) and Type M (Magnitude) errors. *Perspectives on Psychological Science, 9*, 641-651.

Gilbert, D. (2014). Some thoughts on shameless bullies. Available at: http://www.wjh.harvard.edu/~dtg/Bullies.pdf

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197-216.

Hales, A.H. (in press). Does the conclusion follow from the evidence? Recommendations for improving research. *Journal of Experimental Social Psychology*.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61-83.

Hüffmeier, J., Mazei, J., & Schultze, T. (in press). Reconceptualizing replication as a sequence

of different studies: A replication typology. *Journal of Experimental Social Psychology*.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluation of options: A review and theoretical analysis. *Psychological Bulletin, 125(5)*, 576-590.

Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Medicine.* http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0020124

Ioannidis, J. P. A., & Trikalinos T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4*, 245-253.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Jordan, J., Diermeier, D. A., & Galinsky, A. D. (2012). The strategic samaritan: How effectiveness and proximity affect corporate responses to external crises. *Business Ethics Quarterly*, *22*(04), 621–648.

Kahneman, D. (2012). A proposal to deal with questions about priming effects. Retrieved at: http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf

Kahneman, D. (2014). A new etiquette for replication. *Social Psychology*, *45(4)*, 310-311.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences.* Sage.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian,

K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van't Veer, A., Vaughn, L. A., Vranka, M., Wichman, A., Woodzicka, J. A., & Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45(3),* 142–152.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480-498.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., Sartori, G., Dial, C., Sriram, N., Banaji, M. R., & Nosek, B. A. (2014). A comparative investigation of 17 interventions to reduce implicit racial preferences. *Journal of Experimental Psychology: General, 143,* 1765-1785.

Landy, J., & Uhlmann, E.L. (2015). Morality is personal. Invited submission to J. Graham and K. Gray (Eds.) *The atlas of moral psychology.*

Lewandowsky, S., Ecker, U. K. H., Seifert, C., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13,* 106-131.

Lieberman, M.D. (2014). Latitudes of Acceptance. Available online at:

http://edge.org/conversation/latitudes-of-acceptance

Lipsey, M.W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: Sage publications.

Liu, B., & Ditto, P. H. (2013). What dilemma? Moral evaluation shapes factual belief. *Social Psychological and Personality Science, 4,* 316-323.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (in press). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1(2),* 161-175.

Maner, J. K. (in press). Into the wild: Field research can increase both replicability and real world impact. *Journal of Experimental Social Psychology.*

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General, 144*, e1-e15.

Mitchell, J. (2014). On the emptiness of failed replications. Available at: http://wjh.harvard.edu/~jmitchel/writing/failed_science.htm

Molden, D.C., & Higgins, E. T. (2012) Motivated thinking. In K. Holyoak & B. Morrison (Eds.)

The Oxford Handbook of Thinking and Reasoning (pp. 319-335). New York, Psychology Press.

Monin, B., & Merritt, A. (2012). Moral hypocrisy, moral inconsistency, and the struggle for moral integrity. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil*. Washington, DC: American Psychological Association.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review, 112,* 979–999.

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science, 7*, 960-969.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125(6)*, 737.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic vs. analytic cognition. *Psychological Review*, *108*, 291-310.

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin*, 135, 763-784.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23*, 217-243.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137-141.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and

practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615-631.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). DOI: 10.1126/science.aac4716

Pace, K. M., Fediuk, T. A., & Botero, I. C. (2010). The acceptance of responsibility and expressions of regret in organizational apologies after a transgression. *Corporate Communications: An International Journal*, *15*(4), 410–427.

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In P. Shaver & M. Mikulincer (Eds.), *The social psychology of morality: Exploring the causes of good and evil.* New York: APA books.

Pizarro, D.A., Tannenbaum, D., & Uhlmann, E.L. (2012). Mindless, harmless, and blameworthy. *Psychological Inquiry, 23,* 185-188.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712–713.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Sakaluk, J.K. (in press). Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology.*

Schaller, M. (in press). The empirical benefits of conceptual rigor: Systematic articulation of

conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology*.

Schnall, S. (2014a). An experience with a registered replication project. Available at: http://www.psychol.cam.ac.uk/cece/blog#anchor-1

Schnall, S. (2014b). Further thoughts on replications, ceiling effects and bullying. Available at: http://www.psychol.cam.ac.uk/cece/blog

Schnall, S. (2014c). Social media and the crowd-sourcing of social psychology. Available at: http://www.psychol.cam.ac.uk/cece/blog

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470,* 437.

Schooler, J. (2014). Metascience could rescue the 'replication crisis'. *Nature, 515,* 9.

Schwarz, N., & Strack, F. (2014). Does merely going through the same moves make for a ''direct'' replication? Concepts, contexts, and operationalizations. *Social Psychology, 45(3),* 299-311.

Shweder, R. A., & Haidt, J. (1993). The future of moral psychology: Truth, intuition, and the pluralist way. *Psychological Science*, *4*(6), 360–365.

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: *Undisclosed* flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science, 22(11),* 1359-1366.

Simmons, J., Nelson, L., & Simonsohn, U. (2013). Life after p-hacking. Presentation at the Annual Meeting of the Society for Personality and Social Psychology. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2205186http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2205186

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow database on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*, 515-530.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26,* 559-569.

Stanley, D.J. & Spence, J.R. (2014). Expectations for replications: Are you realistic? *Perspectives on Psychological Science*, 9, 305-318.

Stroebe, W. (in press). Are most published social psychological findings false? *Journal of Experimental Social Psychology.*

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9(1)*, 59-71.

Tannenbaum, D., Uhlmann, E.L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology, 47,* 1249-1254.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76(2)*, 105-110.

Uhlmann, E.L., Pizarro, D., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10,* 72-81.

Uhlmann, E.L., Zhu, L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition, 126*, 326-334.

Verhagen, A. J., & Wagenmakers, E.–J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General, 143,* 1457–1475.

Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science, 18*(8), 689–690.

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology, 50,* 149-166.

Wagenmakers, E.-J., Verhagen, A. J., & Ly, A. (in press). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods.*

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J. & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 627-633.

Wilson, T. (2014). Is there a crisis of false negatives in psychology? Available at: http://timwilsonredirect.wordpress.com/2014/06/15/is-there-a-crisis-of-false-negatives-in-psychology/

Yuill, N., Perner, J., Pearson, A., Peerbhoy, D., & Ende, J. (1996). Children's changing understanding of wicked desires: From objective to subjective and moral. *British Journal of Developmental Psychology*, *14*(4), 457–475.

Author Contributions

The first, second, and last authors contributed equally to the project. Eric Luis Uhlmann designed the Pipeline Project and wrote the initial project proposal. Martin Schweinsberg, Nikhil Madan, Michelangelo Vianello, Amy Sommer, Jennifer Jordan, Warren Tierney, Eli Awtrey, and Luke (Lei) Zhu served as project coordinators. Daniel Diermeier, Justin Heinze, Malavika Srinivasan, David Tannenbaum, Eric Luis Uhlmann, and Luke Zhu contributed original studies for replication. Michelangelo Vianello, Jennifer Jordan, Amy Sommer, Eli Awtrey, Eliza Bivolaru, Jason Dana, Clintin P. Davis-Stober, Christilene Du Plessis, Quentin F. Gronau, Andrew C. Hafenbrack, Eko Yi Liao, Alexander Ly, Maarten Marsman, Toshio Murase, Israr Qureshi, Michael Schaerer, Nico Thornley, Christina M. Tworek, Eric-Jan Wagenmakers, and Lynn Wong helped analyze the data. Eli Awtrey, Jennifer Jordan, Amy Sommer, Tabitha Anderson, Christopher W. Bauman, Wendy L. Bedwell, Victoria Brescoll, Andrew Canavan, Jesse J. Chandler, Erik Cheries, Sapna Cheryan, Felix Cheung, Andrei Cimpian, Mark Clark, Diana Cordon, Fiery Cushman, Peter Ditto, Thomas Donahue, Sarah E. Frick, Monica Gamez-Djokic, Rebecca Hofstein Grady, Jesse Graham, Jun Gu, Adam Hahn, Brittany E. Hanson, Nicole J. Hartwich, Kristie Hein, Yoel Inbar, Lily Jiang, Tehlyr Kellogg, Deanna M. Kennedy, Nicole Legate, Timo P. Luoma, Heidi Maibeucher, Peter Meindl, Jennifer Miles, Alexandra Mislin, Daniel Molden, Matt Motyl, George Newman, Hoai Huong Ngo, Harvey Packham, Philip S. Ramsay, Jennifer Lauren Ray, Aaron Sackett, Anne-Laure Sellier, Tatiana Sokolova, Walter Sowden, Daniel Storage, Xiaomin Sun, Christina M. Tworek, Jay Van Bavel, Anthony N. Washburn, Cong Wei, Erik Wetter, and Carlos Wilson carried out the replications. Adam Hahn,

Nicole Hartwich, Timo Luoma, Hoai Huong Ngo, and Sophie-Charlotte Darroux translated study materials from English into the local language. Eric Luis Uhlmann and Martin Schweinsberg wrote the first draft of the paper and numerous authors provided feedback, comments, and revisions. Correspondence concerning this paper should be addressed to Martin Schweinsberg, Boulevard de Constance, 77305 Fontainebleau, France, martin.schweinsberg@insead.edu, or to Eric Luis Uhlmann, INSEAD Organizational Behavior Area, 1 Ayer Rajah Avenue, 138676 Singapore, eric.uhlmann@insead.edu. The Pipeline Project was generously supported by an R&D grant from INSEAD.

Footnote

[1] In a fixed-effects model –i.e. without accounting for between study variability in the computation of the standard error– the bad tipper effect is significantly different from zero in non-USA samples as well.

[2] We thank Raphael Silberzahn for suggesting this approach to implementing PPIRs.

Figures

Figure 1. Original effect sizes (indicated with an X), individual effect sizes for each replication sample (small circles), and meta-analyzed replication effect sizes (large circles). Error bars reflect the 95% confidence interval around the meta-analyzed replication effect size. Note the "Higher Standard" study featured two effects, one of which was originally significant (effect of awarding a small perk to the head of a charity) and one of which was originally a null effect (effect of awarding a small perk to a corporate executive). Note also that the Presumption of Guilt effect was a null finding in the original study (no difference between failure to respond to accusations of wrongdoing and being found guilty).

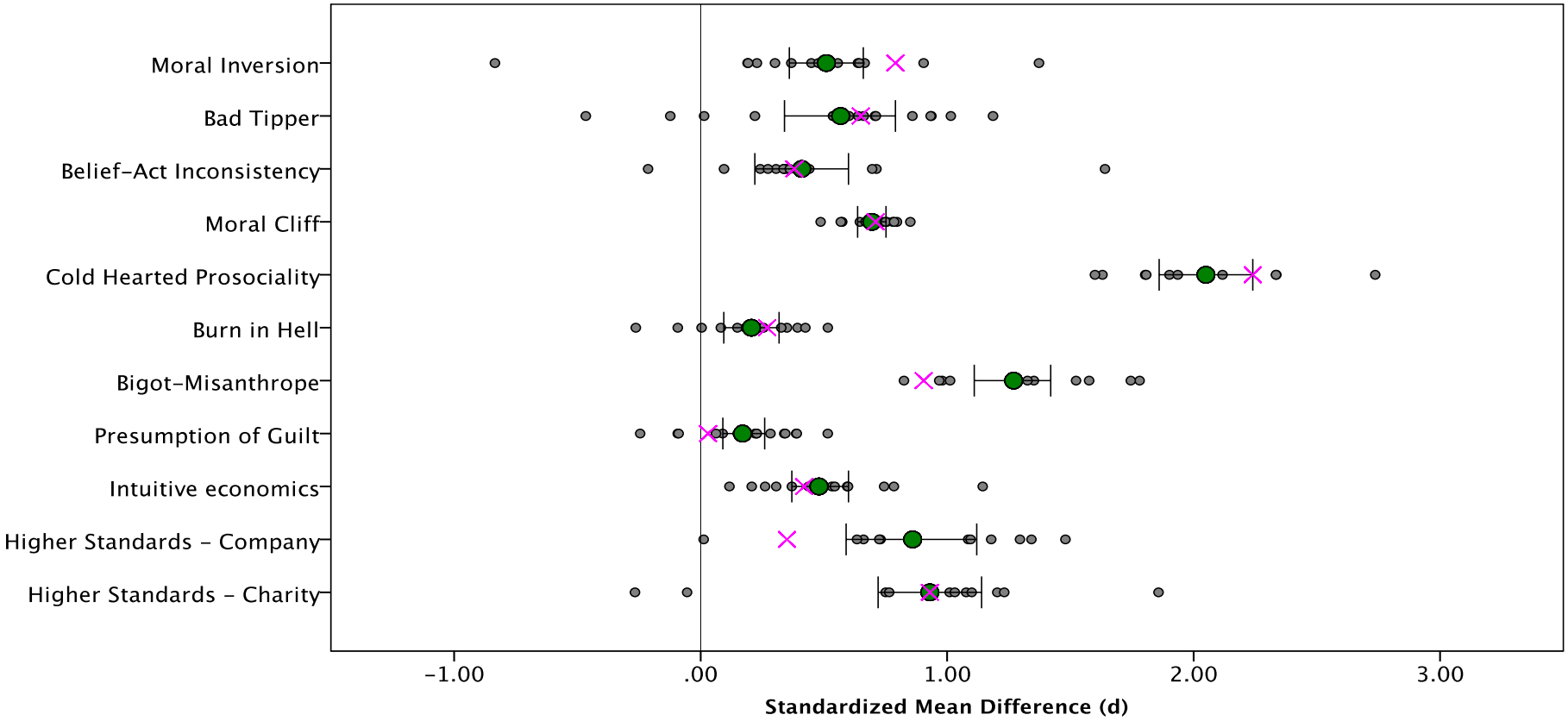

Figure 2. Bayesian inference for the Pipeline Project effects. The y-axis lists each effect and the Bayes factor in favor of or against the default alternative hypothesis for the data from the original study (i.e., BF_{10} and BF_{01}, respectively). The *x*-axis shows the values for the replication Bayes factor where prior distribution under the alternative hypothesis equals the posterior distribution from the original study (i.e., BF_{r0}). For most effects, the replication Bayes factors indicate overwhelming evidence in favor of the alternative hypothesis; hence, the bottom panel provides an enlarged view of a restricted scale. See text for details.

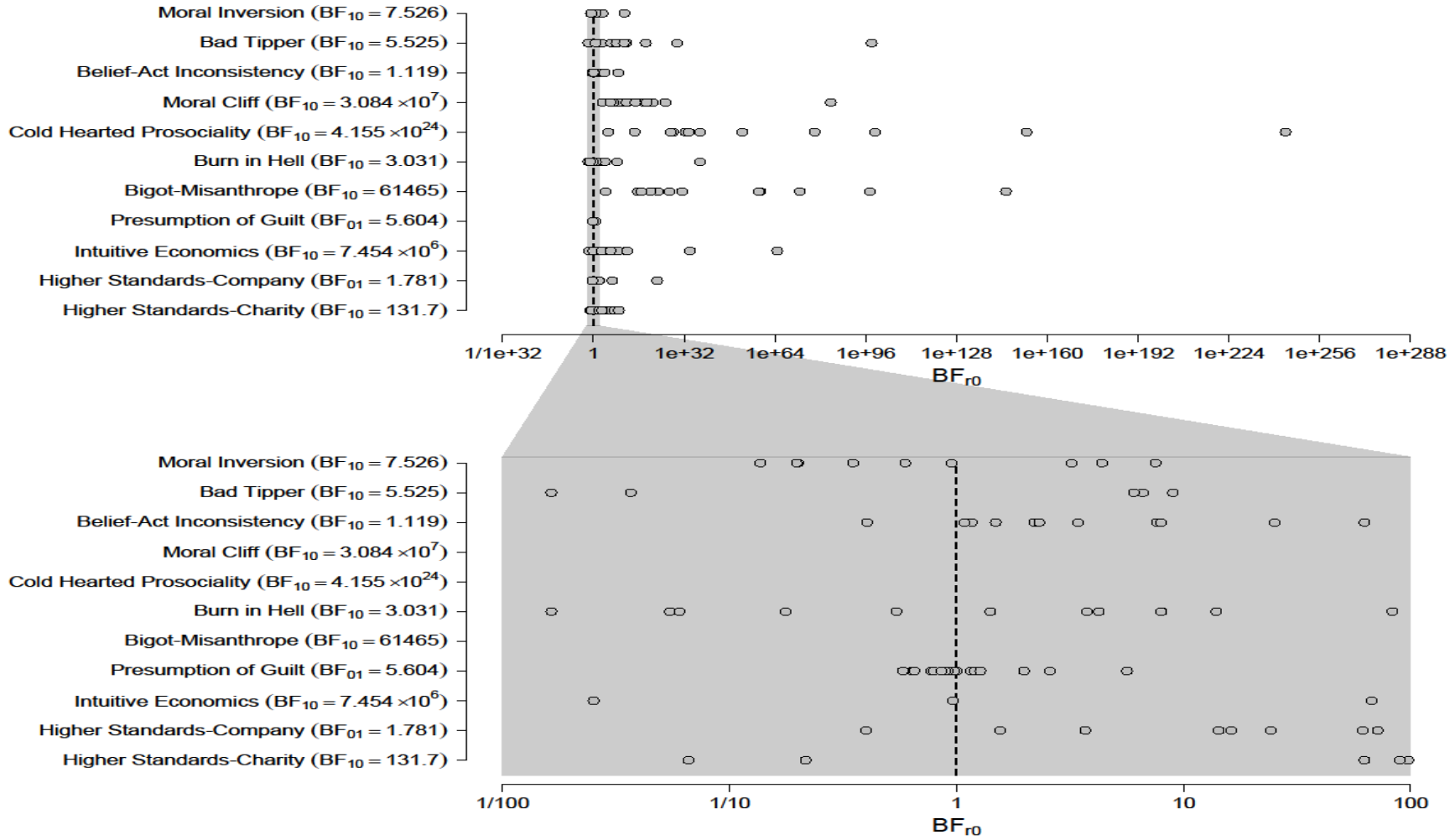

Table 1

Assessment of Replication Results

Effect	Description of original finding	Original and replication effect in same direction	Replication effect significant	Effect significant meta-analyzing the replications and the original study	Original effect inside CI of replication	Small telescopes criterion passed	Overall assessment of replicability
Moral inversion	A company that contributes to charity but then spends even more money promoting the contribution is perceived more negatively than a company that makes no donation at all.	Yes	Yes	Yes	No (rep. < original)	Yes	Successful replication overall. The effect in the replication is smaller than in the original study, but it passes the small telescopes criterion.
Bad tipper	A person who leaves the full tip entirely in pennies is judged more negatively than a person who leaves less money in bills.	Yes	Yes	Yes	Yes	Yes	Replicated robustly in USA with variable results outside the USA.
Belief-act inconsistency	An animal rights activist who is caught hunting is seen as more immoral than a big game hunter.	Yes	Yes	Yes	Yes	Yes	Successful replication
Moral cliff	A company that airbrushes their model to make her skin look perfect is seen as more dishonest than a company that hires a model whose skin already looks perfect.	Yes	Yes	Yes	Yes	Yes	Successful replication
Cold hearted prosociality	A medical researcher who does experiments on animals is seen as engaging in more morally praiseworthy acts than a pet groomer, but also as a worse person	Yes	Yes	Yes	Yes	Yes	Successful replication
Burn in hell	Corporate executives are seen as more likely to burn in hell than vandals.	Yes	Yes	Yes	Yes	Yes	Successful replication

Continued

Table 1

Assessment of Replication Results

Effect	Description of original finding	Original and replication effect in same direction	Replication effect significant	Effect significant meta-analyzing the replications and the original study	Original effect inside CI of replication	Small telescopes criterion passed	Overall assessment of replicability
Bigot-misanthrope	Participants judge a manager who selectively mistreats racial minorities as a more blameworthy person than a manager who mistreats everyone.	Yes	Yes	Yes	No (rep. > original)	Yes	Successful replication
Presumption of guilt	For a company, failing to respond to accusations of misconduct leads to judgments as harsh as being found guilty.	Yes	Yes (original was a null effect)	Yes (original was a null effect)	No (rep. > original)	N/A	Failure to replicate. Original effect was a null effect with a tiny point difference, such that failing to respond to accusations of wrongdoing is just as bad as being investigated and found guilty. However in the replication failing to respond is unexpectedly significantly *worse* than being found guilty, with an effect size over five times as large as in the original study. This cannot be explained by a presumption of guilt as in the original theory.
Intuitive economics	The extent to which people think high taxes are fair is positively correlated with the extent to which they think high taxes are good for the economy.	Yes	Yes	Yes	Yes	Yes	Successful replication

Continued

Table 1

Assessment of Replication Results

Effect	Description of original finding	Original and replication effect in same direction	Replication effect significant	Effect significant meta-analyzing the replications and the original study	Original effect inside CI of replication	Small telescopes criterion passed	Overall assessment of replicability
Higher standards: company condition	It is perceived as acceptable for the top executive at a corporation to receive a small perk.	Yes	Yes (original was a null effect)	Yes (original was a null effect)	No (rep. > original)	N/A	Failure to replicate. The original study found that a small executive perk hurt the reputation of the head of a charity (significant effect) but not a company (null effect). In the replication a small perk hurt both types of executives to an equal degree. The effect of a small perk in the company condition is over two times as large in the replication as in the original study. There is no evidence in the replication that the head of a charity is held to a higher standard.
Higher standards: charity condition	For the leader of a charitable organization, receiving a small perk is seen as moral transgression.	Yes	Yes	Yes	Yes	Yes	

Table 2

Moderators of replication results

Effect	US vs. non-US sample	Study Order	General Population vs. Students	Original Location vs. Different Location
Moral inversion	Not significant	Not significant	Not significant	Significant: Orig > Diff
Bad tipper	Significant: US > non-US	Not significant	Significant: Gen pop > students	Significant: Orig > Diff
Belief-act inconsistency	Not significant	Significant: Late > Early	Not significant	Not significant
Moral cliff	Significant: US > non-US	Not significant	Significant: Gen pop > Students	Significant: Orig > Diff
Cold hearted prosociality	Significant: US > non-US	Not significant	Significant: Gen pop > Students	Not significant
Burn in hell	Significant: US > non-US	Not significant	Significant: Gen pop > Students	Significant: Orig < Diff
Bigot-misanthrope	Significant: US < non-US	Not significant	Significant: Gen pop < Students	Significant: Orig < Diff
Presumption of guilt	Not significant	Not significant	Not significant	Not significant
Intuitive economics	Significant: US > non-US	Not significant	Significant: Gen pop > Students	Not significant
Higher standards: company	Significant: US > non-US	Not significant	Not significant	Significant: Orig > Diff
Higher standards: charity	Not significant	Not significant	Not significant	Not significant